Network Biology Approach to Complex Diseases

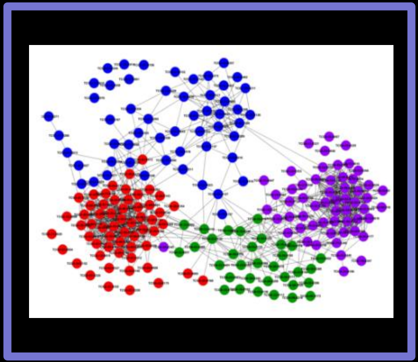
Teresa Przytycka

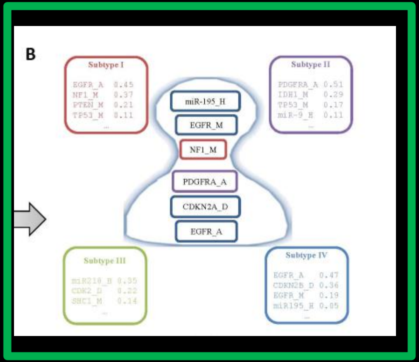
NIH / NLM / NCBI

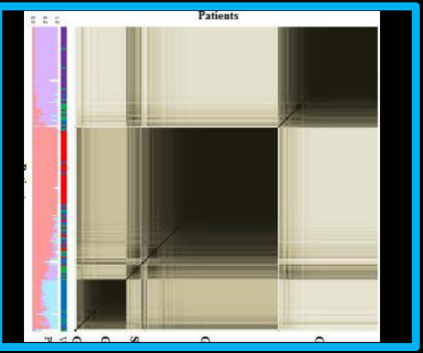
Acknowledgments

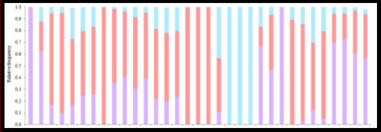
These lectures are based in part the following recent review articles coauthored with DongYeon Cho, YooAh Kim, Jozef Przytycki, Mona Singh, Donna Slonim, and Stefan Wuchty

1. Chapter 5: Network biology approach to complex diseases, Cho DY, Kim YA, Przytycka TM. PLoS Comput Biol. 2012;8(12):
2. Bridging the Gap between Genotype and Phenotype via Network Approaches. Kim YA, Przytycka TM. Front Genet. 2013
3. Modeling information flow in biological networks. Kim YA, Przytycki JH, Wuchty S, Przytycka TM. Phys Biol. 2011
4. Toward the dynamic interactome: it's about time. Przytycka TM, Singh M, Slonim DK.Brief Bioinform. 2010 Jan;11(1):15-29.









Acknowledgments

Przytycka's group

DongYeon Cho
- *Prob. Cancer Model*
- *CNV in fly*

Phuong Dao
- Gene regulation
- Network

Xiangjun Du
- *Non B-DNA e-coli*
- *Non B-DNA human population*

Jan Hoinka
- *Aptamers*

Yoo-Ah Kim
- *Information flow*
- *Module Cover*

Damian Wojtowicz
- *Non-B-DNA, Promoter Structure*
- *Expression noise*

Support: Intramural research program NLM / NIH

Network Biology Approach to Complex Diseases

Organization of the lectures

LECTURE 1. Network Modularity, Genotypic modules

LECTURE 2. Phenotypic / expression based dys-regulated modules: combining expression and genetic data

LECTURE 3. Information flow

LECTURE 4. Disease Heterogeneity

Network Biology Approach to Complex Diseases

LECTURE 1. Network Modularity Genotypic modules

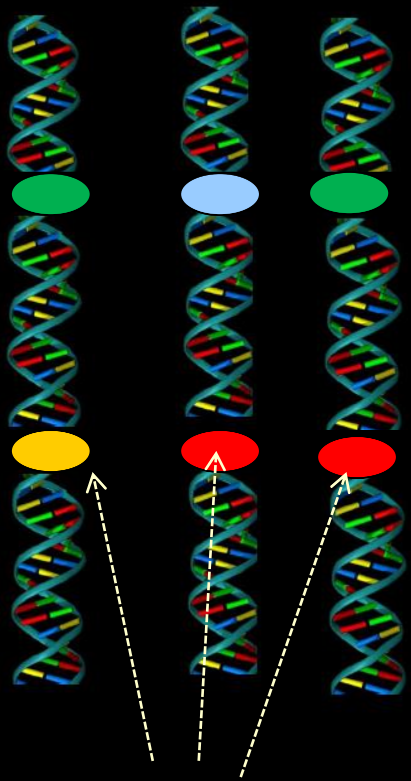
Teresa Przytycka

NIH / NLM / NCBI

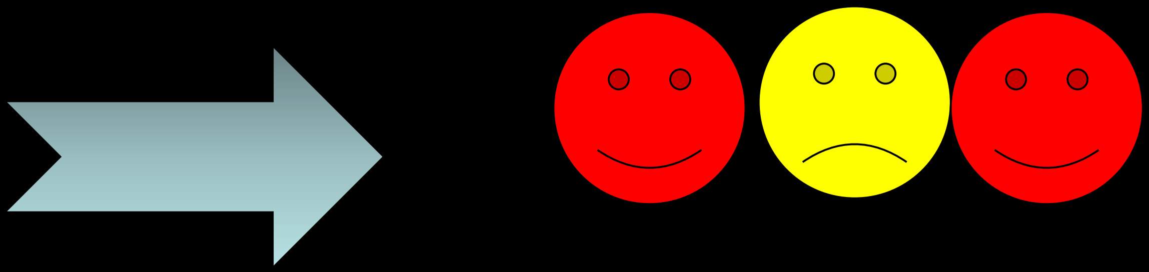
Genotype – Phenotype relation

Individuals (genotype)

Individuals (phenotype)



Genotypic variation



The first step towards linking genotype with phenotype is Genome Wide Association Studies (GWAS) : At which loci the genomic variability correlates with phenotypic variability ?

-- statistical correlation test, corrected for multiple hypothesis testing

Why GWAS is not enough

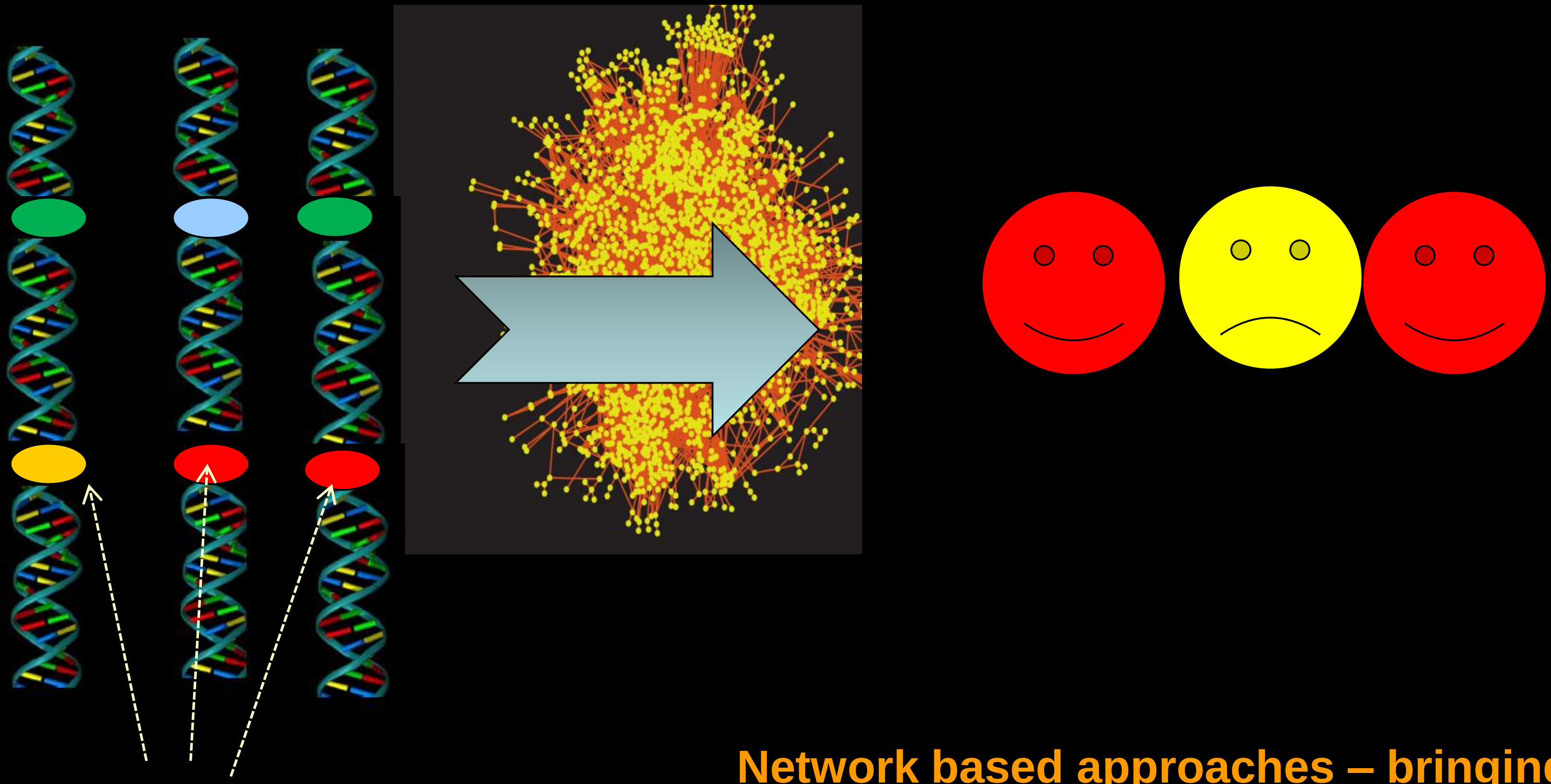
- Complex diseases have multiple causes, which vary from patient to patient

- Individual effects might be small

- Limited Statistical power due to multiple hypothesis testing

- GWAS associations are usually not explanatory

Genotype – Phenotype relation



Network based approaches – bringing knowledge of relation between molecules gained from high throughput experiments

Inferring large scale interaction networks

- Y2H (yeast two hybrid) Protein-protein interactions tests if two proteins can potentially bind

- Co-IP – proteins in these same complex

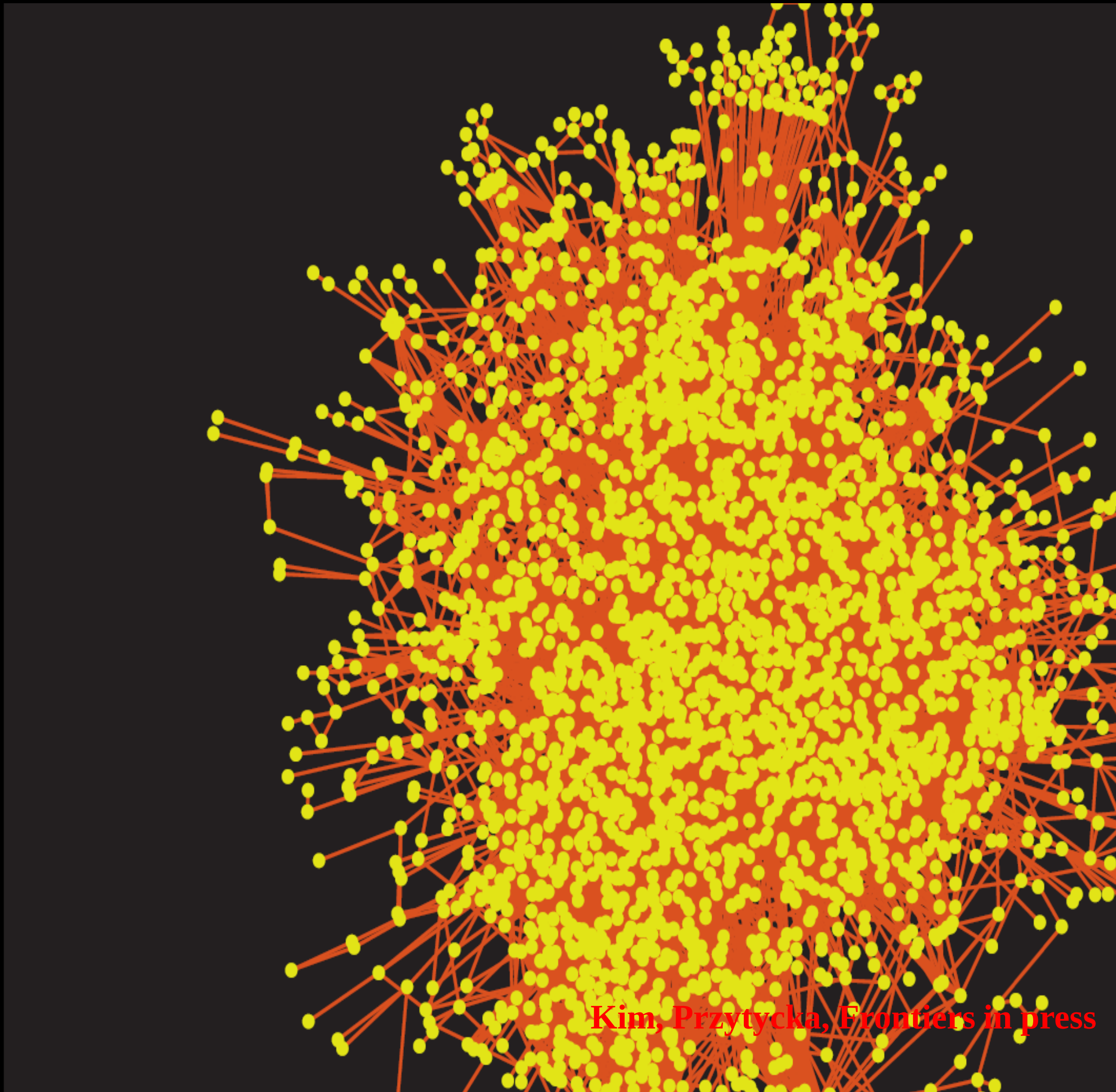
- Genetic interactions – functional relation uncovered when when two genes when perturbed individually have little effect but when perturbed together have a severe effect

- Functional relation inferred based on knowledge of gene function (eg. GO (Gene Ontology) annotation)

- Co-expression networks – Functional interactions predicted from correlation of gene expression over a large number of conditions

- Computational methods based on co-evolution

How to extract information from a high throughput network?



Kim, Przytycka, Frontiers in press

Biological Networks are modular

Module: Group of genes and gene products that work together to preform a specific function

Caveat: We don’t know the function(s) of most genes thus modules need to be predicted from experimentally established relations between genes based on network connectivity.

Exception: Well studied “canonical” pathways

Module identification

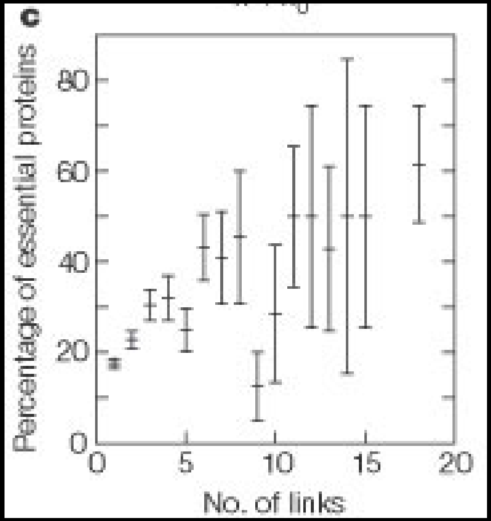
- Huge number of methods – usually as (densely) connected subgraphs based on various connectivity measures
- Our focus –modules/subnetworks related to disease

Enrichment analysis

- Given as module, or other set of genes we ask if it contains more genes from a particular category/function than expected by chance
- Sources of functional annotation GO terms, DAVID (has also dieses association terms)
- Number of software tools Panther, DAVID,

Case study underlining importance of thinking in terms of modules

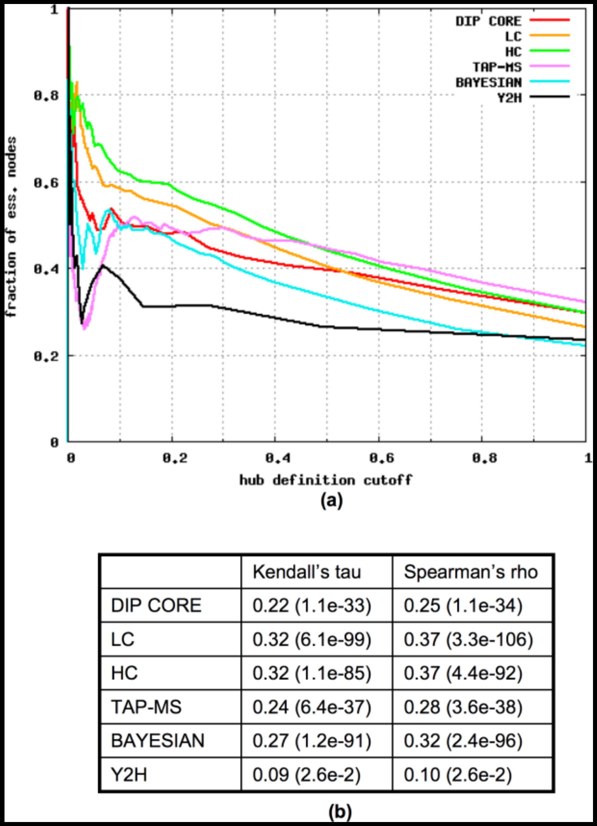
Why are hubs enriched in essential proteins?



H.Jeong et.al. Nature (2001) 411:41-42

Enrichment of hubs in essential nodes

The enrichment depends on network type



(a)

	Kendall's tau	Spearman's rho
DIP CORE	0.22 (1.1e-33)	0.25 (1.1e-34)
LC	0.32 (6.1e-99)	0.37 (3.3e-106)
HC	0.32 (1.1e-85)	0.37 (4.4e-92)
TAP-MS	0.24 (6.4e-37)	0.28 (3.6e-38)
BAYESIAN	0.27 (1.2e-91)	0.32 (2.4e-96)
Y2H	0.09 (2.6e-2)	0.10 (2.6e-2)

(b)

Zotenko, Mestre, O'Leary, Przytycka. PloS CB 2008
(highlighted in Nature Genetics Rev, Sept 2008)

Why are hubs enriched in essential proteins?

- **The Centrality Hypothesis:** If removal of a node disrupts the "communication" between pairs of other nodes in the network, then the corresponding protein is likely to be essential (Jeong et al., Nature 2001)

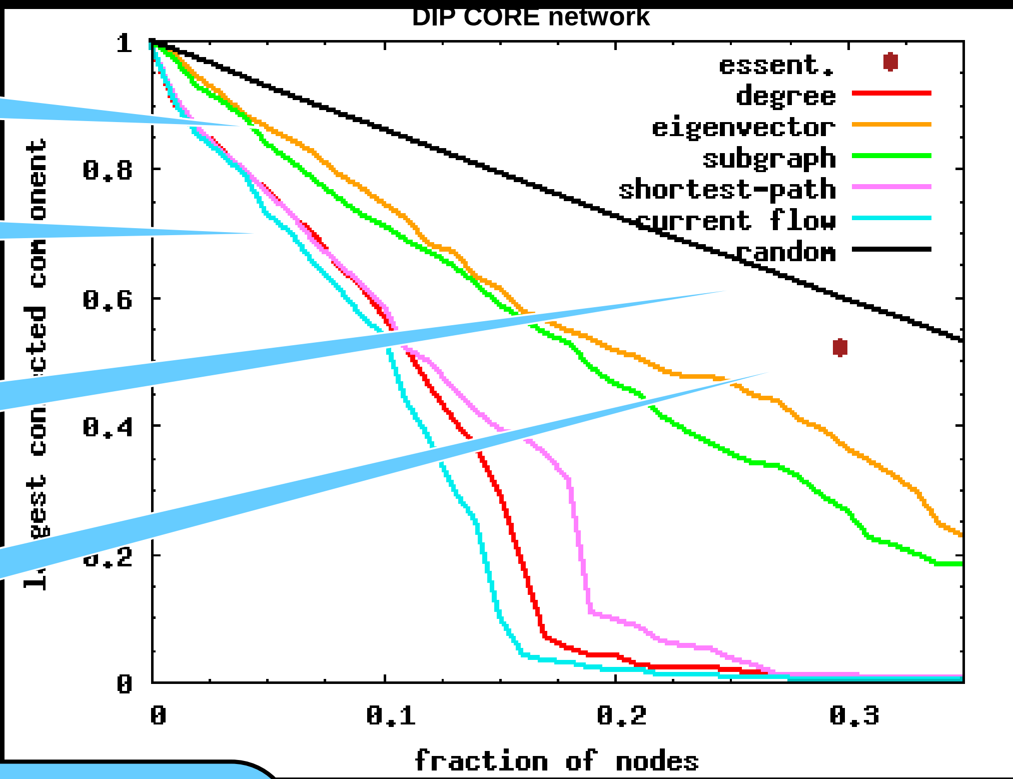
- **The Essential PPIs Hypothesis:** All interactions are essential with uniform probability. High degree nodes are essential because they participate in many interactions and thus, with high probability, are adjacent to an essential interaction (He et al., PLoS Genetics 2006)

Network Centrality Indices

A centrality index assigns a centrality value to every node in the network which quantifies its topological prominence.

- Local indices (how important is the node locally)
  - Degree Centrality (DC)
    - c(v) is the number of neighbors
  - Subgraph Centrality (SC)
    - c(v) is the number of closed walks that start and terminate at v
- Betweenness indices (how important is the node globally)
  - Shortest-Path Betweenness Centrality (SPBC)
    - c(v) is the fraction of shortest paths that pass through v
  - Current Flow Betweenes Centrality (CFBC)
    - c(v) extends the shortest-path betweenness values by taking into account other paths and allowing weights

How destructive to network integrity is removal of central nodes



local indices

betweenness indices

random proteins

essential proteins

Network Integrity Measures

- fraction of nodes in the largest connected component
- increase in the average shortest path
- decrease in the number of edge-disjoint paths

Zotenko, Mestre, O'Leary, Przytycka. PloS CB 2008
(highlighted in Nature Genetics Rev, Sept 2008)

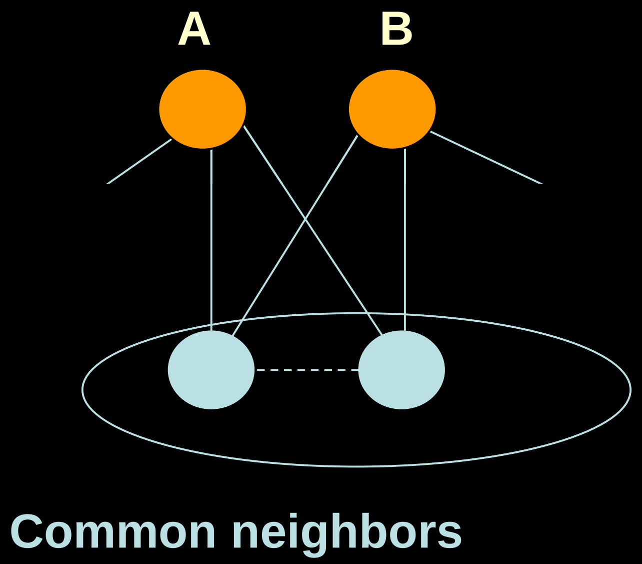
Why are hubs enriched in essential proteins?

- **The Centrality Hypothesis:** If removal of a node disrupts the "communication" between pairs of other nodes in the network, then the corresponding protein is likely to be essential (Jeong et al., Nature 2001)

- **The Essential PPIs Hypothesis:** All interactions are essential with uniform probability. High degree nodes are essential because they participate in many interactions and thus, with high probability, are adjacent to an essential interaction (He et al., PLoS Genetics 2006)

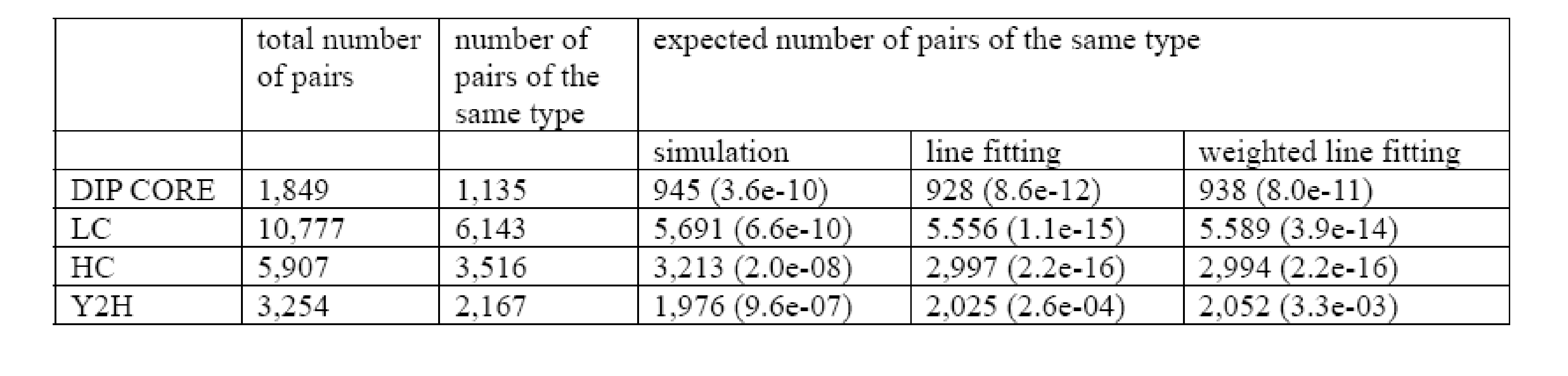
- **Our result:** Neither of the above is true. Alternative view is proposed.

Rejecting Essential PPIs Hypothesis:



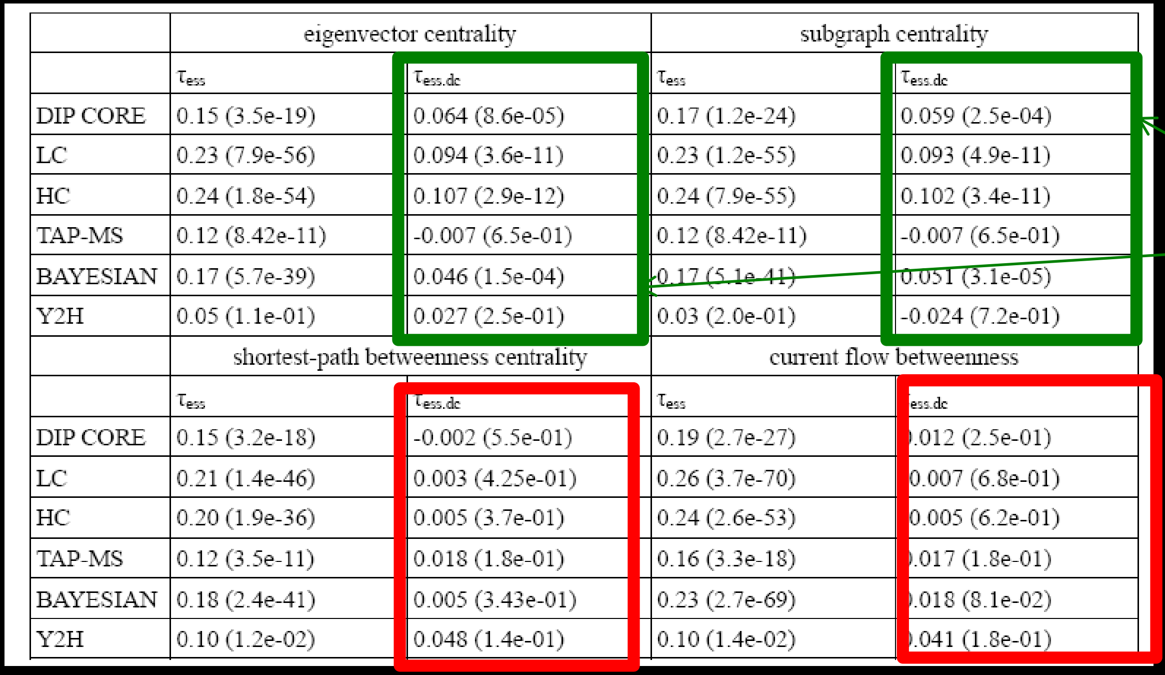
According to the essential interaction hypothesis, essentiality of A should be independent of essentiality of B.

The independence of such pairs was rejected with high probability



	total number of pairs	number of pairs of the same type	expected number of pairs of the same type		
			simulation	line fitting	weighted line fitting
DIP CORE	1,849	1,135	945 (3.6e-10)	928 (8.6e-12)	938 (8.0e-11)
LC	10,777	6,143	5,691 (6.6e-10)	5.556 (1.1e-15)	5.589 (3.9e-14)
HC	5,907	3,516	3,213 (2.0e-08)	2,997 (2.2e-16)	2,994 (2.2e-16)
Y2H	3,254	2,167	1,976 (9.6e-07)	2,025 (2.6e-04)	2,052 (3.3e-03)

Correlation of global centrality measures with essentiality is not statistically significant when correcting for correlation with vertex degree



	eigenvector centrality		subgraph centrality	
	τ_{ess}	$\tau_{ess.dc}$	τ_{ess}	$\tau_{ess.dc}$
DIP CORE	0.15 (3.5e-19)	0.064 (8.6e-05)	0.17 (1.2e-24)	0.059 (2.5e-04)
LC	0.23 (7.9e-56)	0.094 (3.6e-11)	0.23 (1.2e-55)	0.093 (4.9e-11)
HC	0.24 (1.8e-54)	0.107 (2.9e-12)	0.24 (7.9e-55)	0.102 (3.4e-11)
TAP-MS	0.12 (8.42e-11)	-0.007 (6.5e-01)	0.12 (8.42e-11)	-0.007 (6.5e-01)
BAYESIAN	0.17 (5.7e-39)	0.046 (1.5e-04)	0.17 (5.1e-41)	0.051 (3.1e-05)
Y2H	0.05 (1.1e-01)	0.027 (2.5e-01)	0.03 (2.0e-01)	-0.024 (7.2e-01)
	shortest-path betweenness centrality		current flow betweenness	
	τ_{ess}	$\tau_{ess.dc}$	τ_{ess}	$_{ess.dc}$
DIP CORE	0.15 (3.2e-18)	-0.002 (5.5e-01)	0.19 (2.7e-27)	.012 (2.5e-01)
LC	0.21 (1.4e-46)	0.003 (4.25e-01)	0.26 (3.7e-70)	0.007 (6.8e-01)
HC	0.20 (1.9e-36)	0.005 (3.7e-01)	0.24 (2.6e-53)	0.005 (6.2e-01)
TAP-MS	0.12 (3.5e-11)	0.018 (1.8e-01)	0.16 (3.3e-18)	.017 (1.8e-01)
BAYESIAN	0.18 (2.4e-41)	0.005 (3.43e-01)	0.23 (2.7e-69)	.018 (8.1e-02)
Y2H	0.10 (1.2e-02)	0.048 (1.4e-01)	0.10 (1.4e-02)	.041 (1.8e-01)

Partial correlation controlled for degree

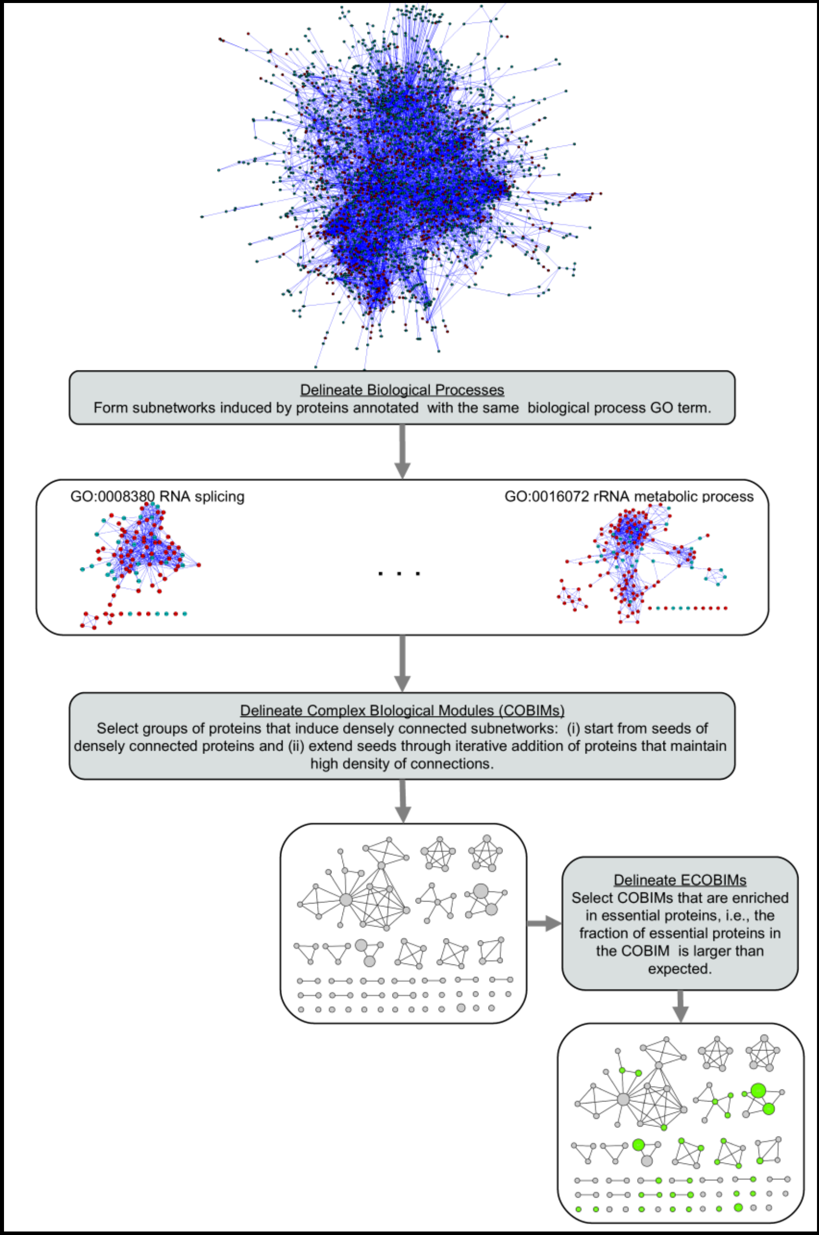
Zotenko, Mestre, O'Leary, Przytycka. PloS CB 2008 (highlighted in Nature Genetics Rev, Sept 2008)

Modularity of Response

Essentiality of hubs is explained by membership in Essential COmplex Biological Modules (ECOBIMs)

Complex Biological Module (COBIM) is a group of proteins that:
- share a biological function (Biological Module)
- interact extensively with each other (Complex)

COBIMs are clearly partitioned into two classes:
- enriched in essential proteins (ECOBIMs)
- depleted of essential proteins



Delineate Biological Processes
Form subnetworks induced by proteins annotated with the same biological process GO term.

Delineate Complex Biological Modules (COBIMs)
Select groups of proteins that induce densely connected subnetworks: (i) start from seeds of densely connected proteins and (ii) extend seeds through iterative addition of proteins that maintain high density of connections.

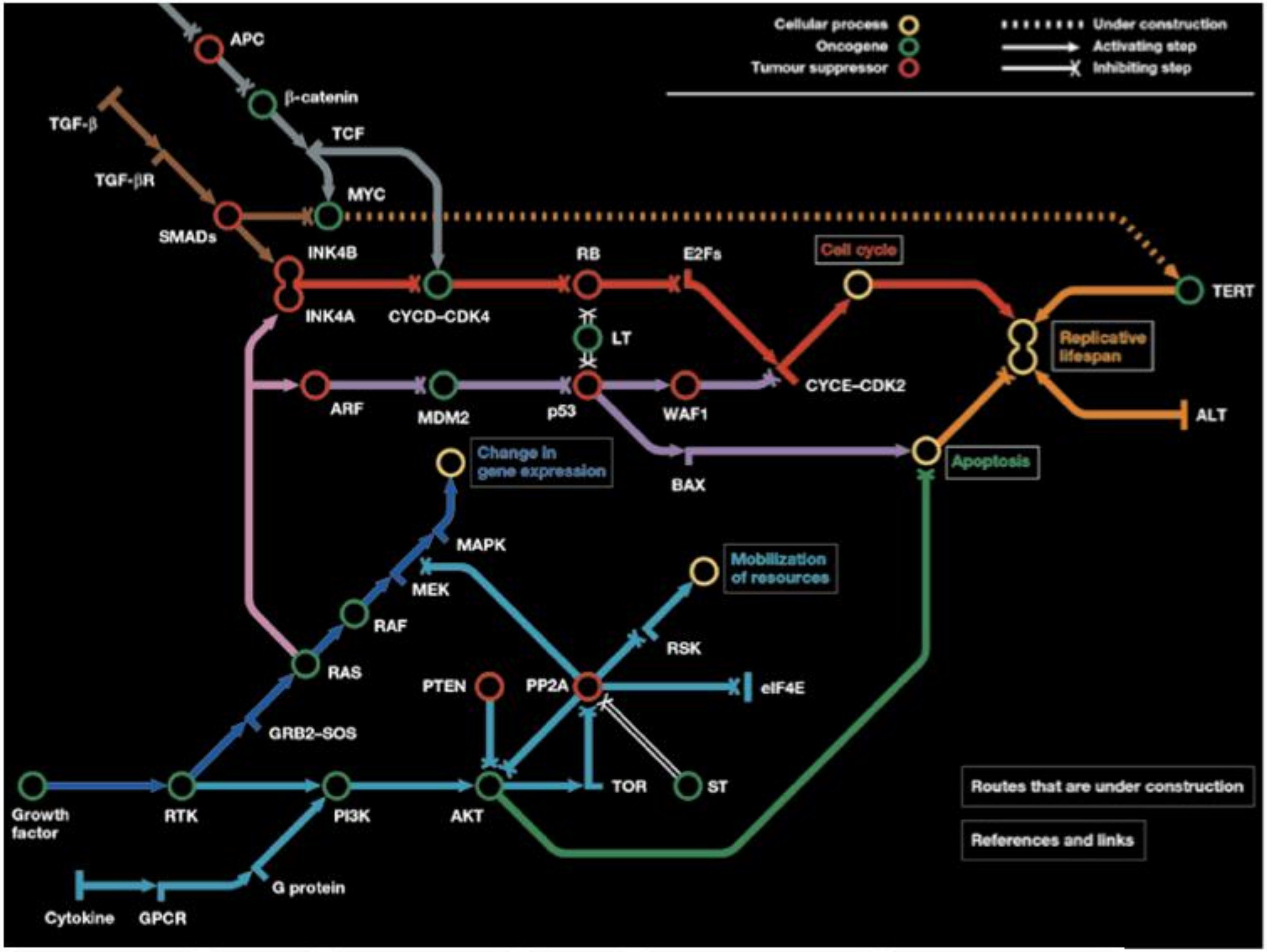
Delineate ECOBIMs
Select COBIMs that are enriched in essential proteins, i.e., the fraction of essential proteins in the COBIM is larger than expected.

Network Modularity in the context of diseases:

Dys-regulated pathways hypothesis

- In complex diseases different genetic / epigenetic causes dysregulate the same molecular pathway(s) / module(s) which therefore leading to similar disease phenotype
- Example – cancer is dysregulation of cell proliferation pathway

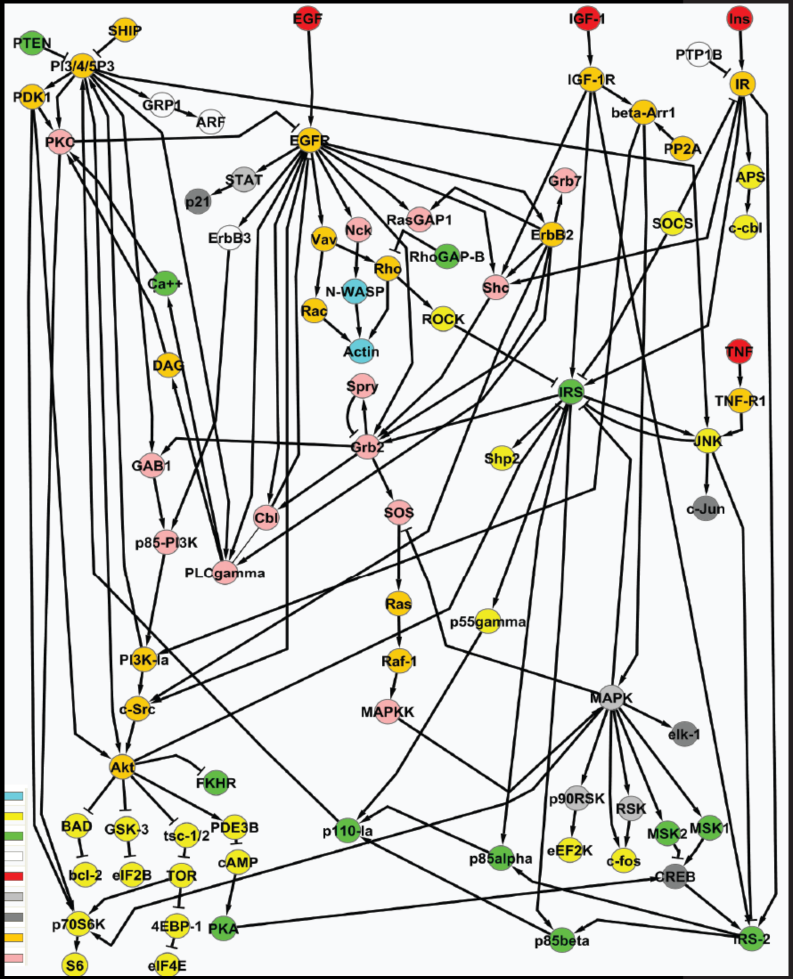
(but we hope to be able to identify more specific pathways)

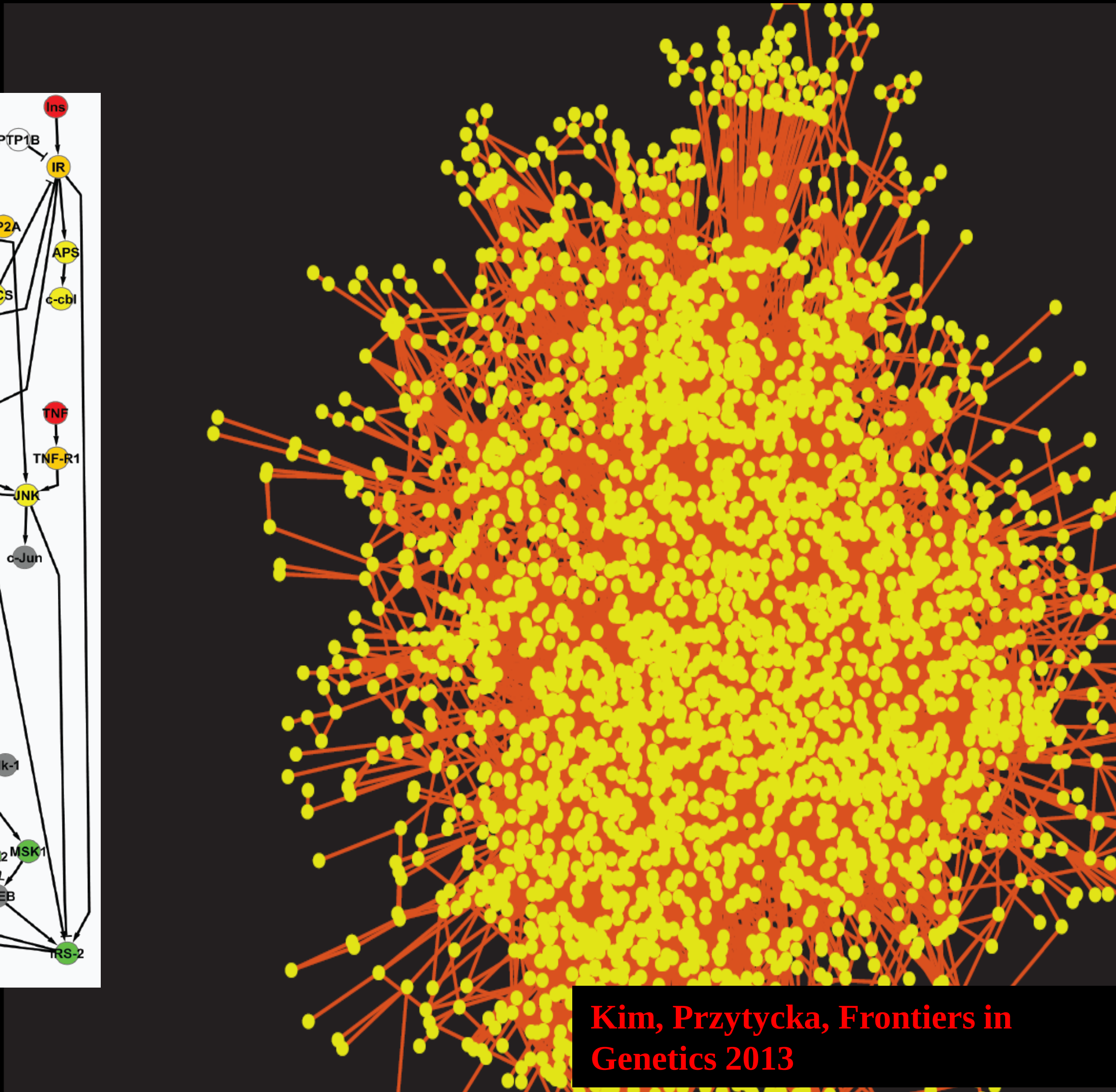


Replicative lifespan | Apoptosis | Proliferative signals | Cell cycle | Mobilization of resources | Routes under construction

Subway map designed by Claudia Bentley.
Web design by Nick Allin.
Edited by Cath Brooksbank and Sandra Clark.

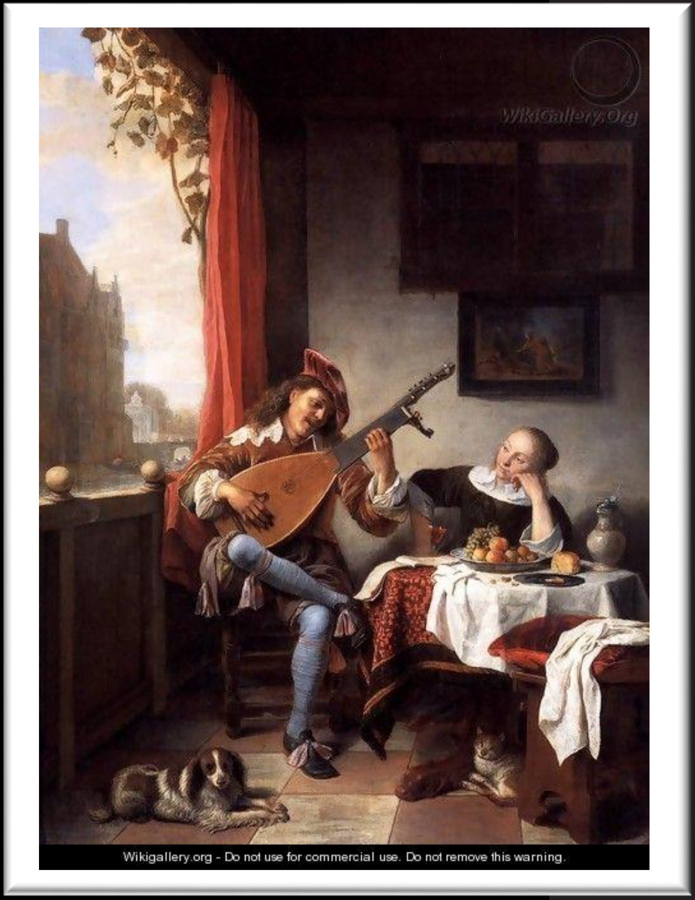
High throughput versus networks derived by small scale experiments



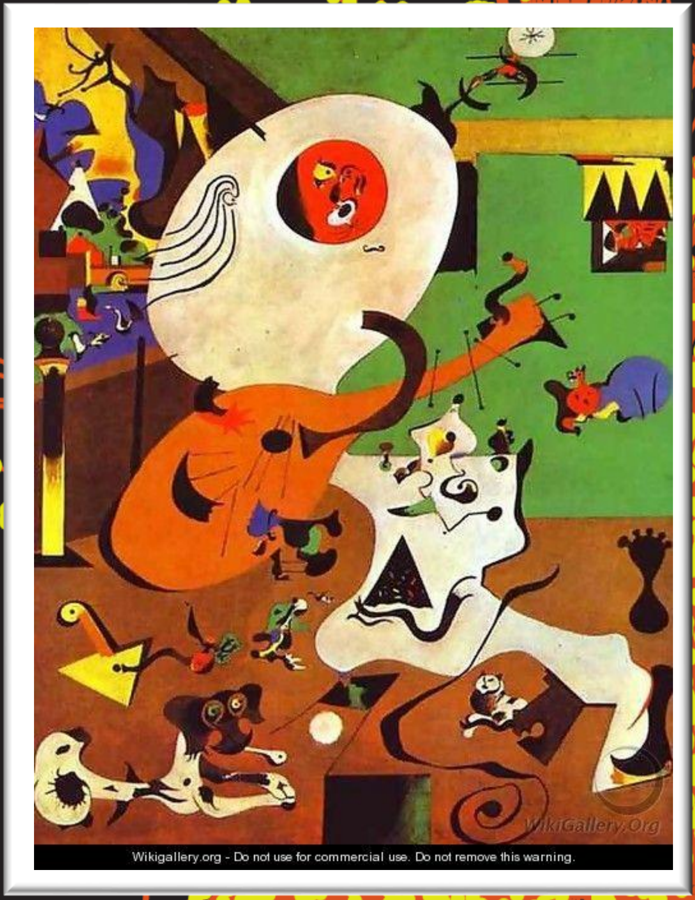


Kim, Przytycka, Frontiers in Genetics 2013

High throughput network versus “the true” network



The Lute Player, Hendrick Maertensz Sorgh (1610-1670), Rijksmuseum, Amsterdam
(*public domain*)



Dutch Interior 1, Joan Miro' (1893–1983)
Museum of Modern Art, New York

Kim, Przytycka, Frontiers 2013

Properties of modules

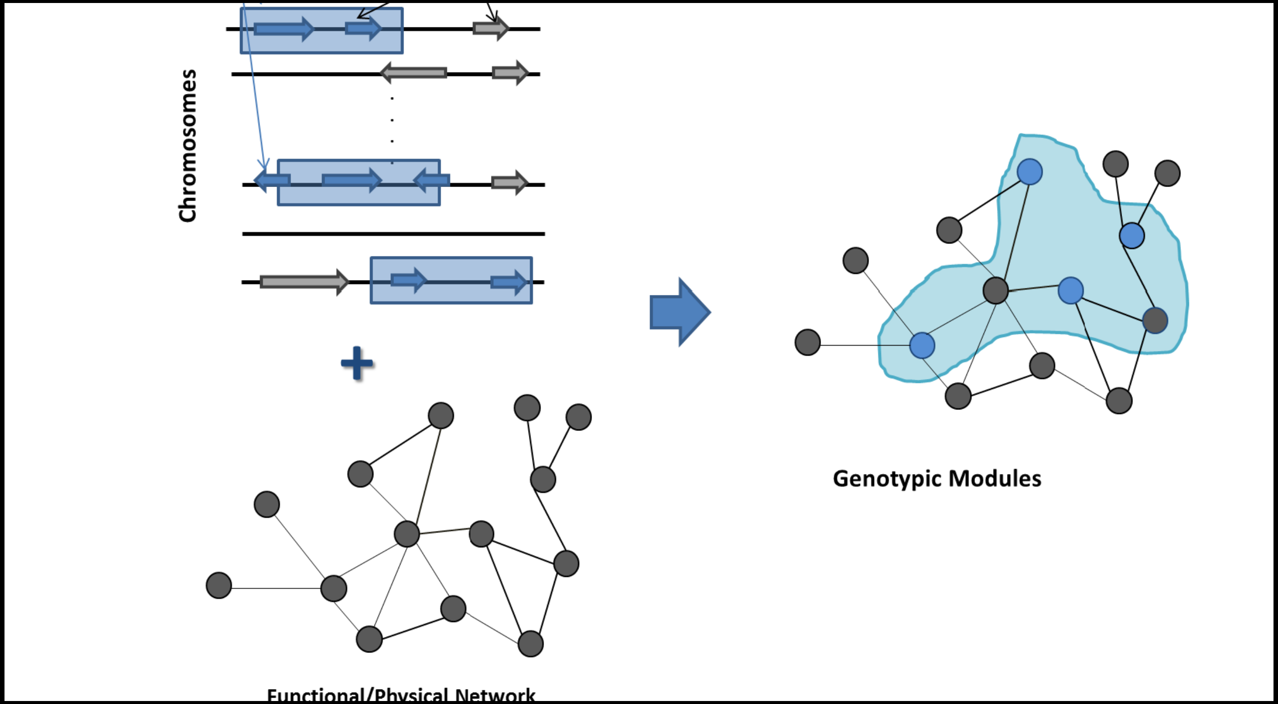
- Organized in hierarchical fashion
- Composed of many different types of molecules – diverse interaction types
- Functions of modules are (more or less) discrete entities and arise as a result of interactions among its components
- Overlapping & Dynamic

  Individual genes can belong to several modules either simultaneously or at various time points

Three different angles in uncovering disease associated modules

- Genotypic modules
  - Modules enriched in causative mutations
- Phenotypic modules
  - Modules enriched over/under expressed genes
- Pathways connecting genotype and phenotype
  - Pathways connecting mutations to abnormally expressed genes

Genotypic modules



Searching for genotypic modules:

- identification of genes/genomic regions that are frequently altered in a disease of interest
- mapping the genes residing in the altered regions to a network
- modules or subnetworks enriched with the altered genes are identified

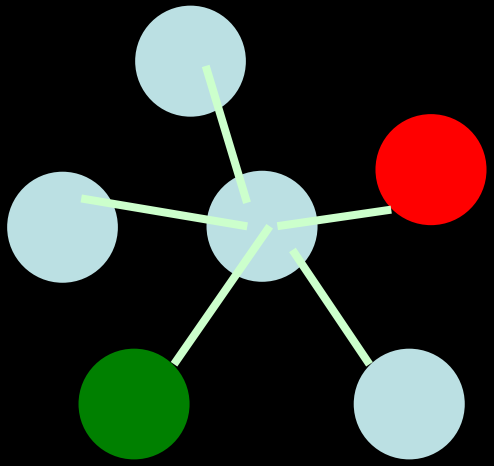
Individual approaches differ in the way this last step is preformed

Example 1: HOTNET – identification of subnetworks using diffusion process

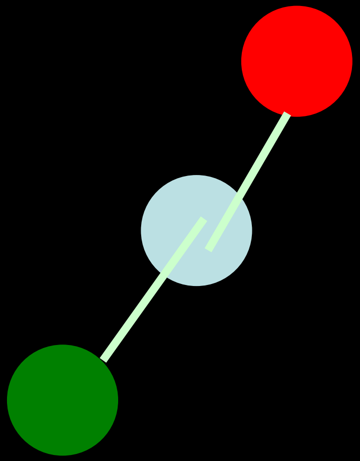
Vandin, F., E. Upfal, and B.J. Raphael, *Algorithms for detecting significantly mutated pathways in cancer.* J Comput Biol, 2011. 18(3): p. 507-22.

- Using protein interaction network construct weighted influence graph where influence between a pair of genes in computed using diffusion process
- Identify significant subnetworks of fixed size covering maximum number of disease cases
- Assessing significance - permutation test

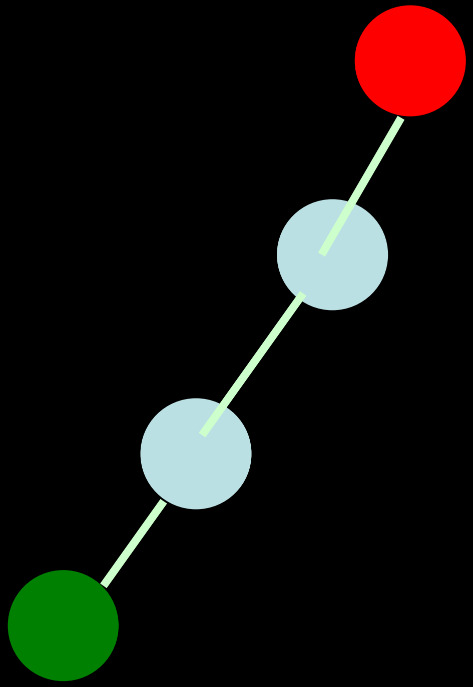
Constructing influence graph - Heat diffusion with heat loss along the edges (related to current flow but current flow has no current loss). The Influence graph – contains all pairs of nodes with influence above a threshold



Lower influence



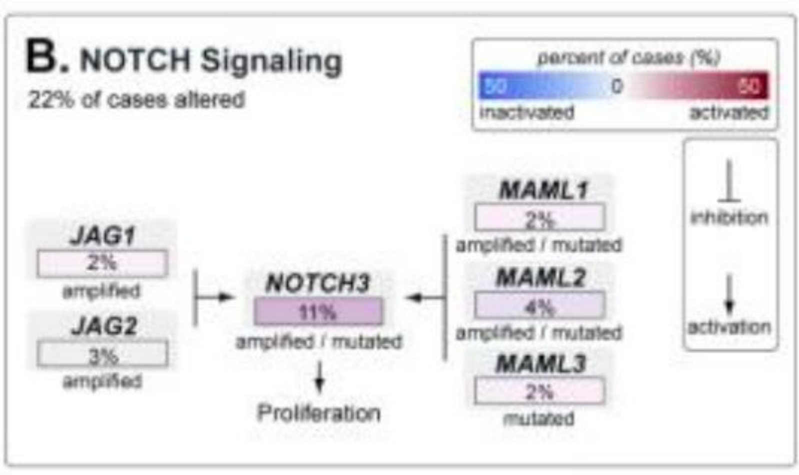
Higher influence



Lower influence

Selecting significant subnetworks – connected cover – (more at a later lecture)

Hotnet identifies significant mutation in Notch signaling pathway in Ovarian cancer



Integrated genomic analyses of ovarian carcinoma. Nature, 2011. 474(7353): p. 609-15.

Example 2: NETBAG

Gilman, S.R., et al., *Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses.* Neuron, 2011. 70(5): p. 898-907

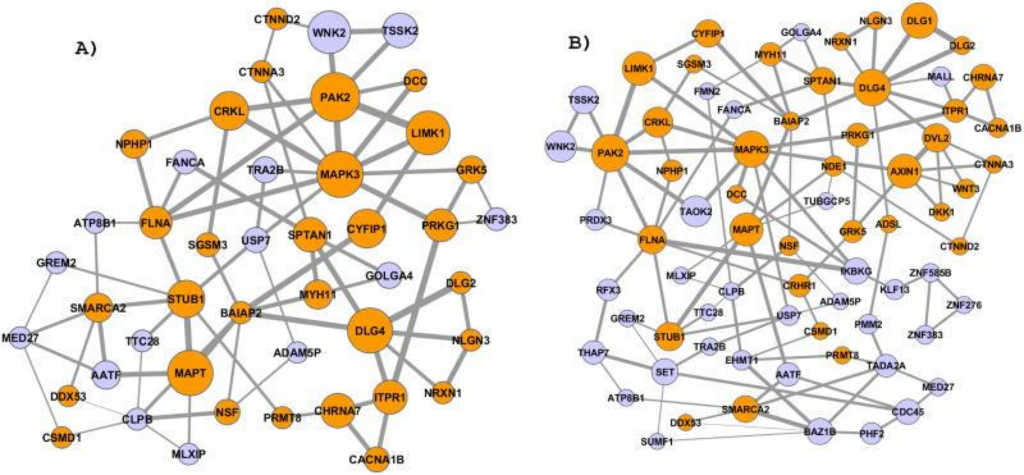
- Naïve Bayes approach is used to construct background network (the network is build form GO annotations, protein-protein interactions, sequence homology etc. and using data from Feldman et al. 2008 as training) edges are assigned the likelihood odd ratio for contributing to the same genetic phenotype

- Genes with CNV were then mapped to the likelihood network and connected clusters of such genes were identified.

- A greedy growth algorithm was used to find the cluster with maximal score

- The significance of a cluster score was estimated by the distribution of maximal scores for clusters obtained from randomized data.

Gene cluster found using NETBAG analysis of rare de-novo copy number variations in autism



In the figure genes (nodes) with known functions in the brain and nervous systems are colored in orange (node size - importance to the overall cluster score; edges - likelihood of shared phenotype)

Modules are enriched in synapse development, axon targeting, etc.

Summary: Advantages of pathway-centric approach

GWAS shortcomings

- Complex diseases have multiple causes, which vary from patient to patient
- Individual effects might be small
- Loss of statistical power due to multiple hypothesis testing
- GWAS associations are usually not explanatory

pathway-centric approaches

- Despite multiple causes dys-regulated pathways might be the same in many disease cases
- Composite effect from whole pathway is likely to be significantly stronger
- Smaller number of tests
- Networks are more often explanatory
- Function of a module is easier to interpret than function of a gene